Designing Experiments to Understand the Variability in Biochemical Reaction Networks

Jakob Ruess, Andreas Milias-Argeitis and John Lygeros [*]

Automatic Control Laboratory, CH-8092 Zurich, ETH Zurich, Switzerland

Abstract

Exploiting the information provided by the molecular noise of a biological process has proven to be valuable in extracting knowledge about the underlying kinetic parameters and sources of variability from single cell measurements. However, quantifying this additional information a priori, to decide whether a single cell experiment might be beneficial, is currently only possibly in very simple systems where either the chemical master equation is computationally tractable or a Gaussian approximation is appropriate. Here we show how the information provided by distribution measurements can be approximated from the first four moments of the underlying process. The derived formulas are generally valid for any stochastic kinetic model including models that comprise both intrinsic and extrinsic noise. This allows us to propose an optimal experimental design framework for heterogeneous cell populations which we employ to compare the utility of dual reporter and perturbation experiments for separating extrinsic and intrinsic noise in a simple model of gene expression. Subsequently, we compare the information content of different experiments which have been performed in an engineered light-switch gene expression system in yeast and show that well chosen gene induction patterns may allow one to identify features of the system which remain hidden in unplanned experiments.

Introduction

Quantitative studies of biological systems with mathematical models strongly depend on an appropriate characterization of the underlying system, that is on good knowledge about the underlying mechanisms and kinetic parameters. While extracting such knowledge from averaged cell population data is common practice, it has only recently been realized that also the molecular noise observed in single cell measurements may be a rich source of information about the parameters of stochastic kinetic models [1, 2, 3, 4]. Mathematically, the information provided by single cell measurements can be quantified with the Fisher information. To compute the Fisher information for stochastic kinetic models one has to solve the chemical master equation which governs the time evolution of the probability distribution of the underlying process [5]. This is, however, only possible in the simplest cases and even approximation techniques either remain limited to very small systems [6] or are based on strong assumptions [7, 8] which are rarely fulfilled in real applications. Consequently, experiments are usually designed based on the intuition of the experimenter, rather than on information theoretic criteria.

A second difficulty in the analysis and design of single cell experiments is that stochastic kinetic models are usually based on the assumption that the same process governs the evolution in all cells of the population. However, in reality cells are different, due to different cell size, local growth conditions, expression capacity or several other factors [9, 10]. This so-called extrinsic variability [11, 12, 13] makes the process evolve slightly different in each cell, often leading to additional variability in the response of the cell population. Indeed, in many instances noise resulting from such extrinsic variability has been reported to dominate molecular noise [14, 15, 16]. In such situations methods that assume a homogeneous cell population and attribute all the observed variability to molecular noise may lead to biased results. Extrinsic variability does not necessarily refer to "extrinsic to the cell" but extrinsic to the model developed for a specific process. Some of this variability may be static for the time scales of interest. For example, the number of mitochondria affecting translation [3] changes much more slowly than species in signaling and transcription cascades which form the heart of the model. In this case the number of mitochondria can be taken as random but constant in time for the purposes of the model. Some of the extrinsic variability, however, may be due to species which are not included in the model but affect reaction rates and evolve on time scales comparable to those of the reactions of interest [11, 17, 18], for instance global regulators affecting transcription of genes of interest. In theory one should include such species in the model but this is often not practical since it would lead to models of intractable size. A convenient modeling abstraction may then be to include an assumed stochastic process for some of the rates, to serve as a rudimentary abstraction of the complex mechanisms governing the fluctuations of the reaction rates [10].

In this paper we develop a framework for optimally designing single cell distribution experiments which is applicable in the presence of possibly time-varying extrinsic variability and scalable to systems of realistic size. To this end we first demonstrate on systems where one reaction rate is governed by a stochastic differential equation how equations describing the time evolution of the moments of the probability distribution can be derived

[*] lygeros@control.ee.ethz.ch

for systems which are influenced by time-varying extrinsic variability. We then show how the Fisher information can be approximated from the first four moments without the need of any assumption other than a sufficiently large measured cell population. Finally, we embed the approximated Fisher information into an optimization algorithm which returns the most informative experiment for a specified set of model parameters. This allows us to design optimal experiments for identifying specific features of the system. We demonstrate this by comparing dual reporter and perturbation experiments in a simple model of gene expression where the mRNA production rate is varying according to a stochastic differential equation. Finally, we study the variability in an engineered light-switch gene expression system in yeast. We use our methodology to evaluate the experiments that were performed in [19] and show that they strongly differ in the information provided about the unknown parameters. Further, we show that an experiment found by employing our optimal experimental design procedure would lead to far more information than any of the experiments reported in [19].

Methods

Moment equations for reaction systems in fluctuating random environments The time evolution of the probability distribution of stochastic kinetic models is governed by the chemical master equation (CME). If extrinsic variability is present in a population, the distribution for each cell can be described by a CME which is conditioned on the realizations of the extrinsic variables in each cell. In [3] it was shown how population moments can be computed from this conditional CME under the assumption that the extrinsic variables are constant in time.

More generally, assume now that extrinsic variability can be modeled by a reaction parameter a_t governed by a stochastic differential equation of the form

$$da_t = r(\mu_a - a_t)dt + s\sqrt{a_t}dW_t, \tag{1}$$

where W_t is a standard Brownian motion. This process fluctuates around its mean μ_a, where the mean reversion speed r gives the autocorrelation time of the process and thereby determines the time scale of the rate fluctuation. The noise coefficient s determines the size of the deviations from the mean, whereas the term $\sqrt{a_t}$ prevents the process from taking negative values and is in accordance with the frequently used Langevin approximation for chemical reaction networks [8]. Note that this formulation includes constant extrinsic variability since for $r = s = 0$ a_t is constant and distributed according to its initial distribution.

The system which jointly describes the time evolution of the species and the extrinsic variable is a stochastic hybrid system [20]. The time evolution of the moments can be computed as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbb{E}\left[\psi(a_t, x(t))\right] = \mathbb{E}\left[(L\psi)(a_t, x(t))\right], \tag{2}$$

where $x(t) = [x_1(t) \cdots x_m(t)]$ is a vector containing the molecule counts of the m species, $\psi : \mathbb{R} \times \mathbb{R}^m \longrightarrow \mathbb{R}$ is chosen such that the left hand side gives the derivative of the desired moment and L is the extended generator of the stochastic hybrid system, given by

$$\begin{aligned}(L\psi)(a,x) :=& \frac{\partial\psi(a,x)}{\partial a} \cdot f(a) + \frac{1}{2}\frac{\partial^2\psi(a,x)}{\partial a^2} \cdot s^2 a \\ &+ \sum_{i=1}^{K} \left(\psi(a, x + \nu_i) - \psi(a,x)\right) a_i(a,x),\end{aligned}$$

where $f(a) = r(\mu_a - a)$, ν_i are the stoichiometric transition vectors and $a_i(a,x)$ the propensities of the K reactions. Note that the resulting system of moment equations may be non-closed in the sense that the time evolution of the moments of any order depends on moments of higher order. In such cases the moments cannot be computed exactly and approximation techniques have to be used [21, 22, 23].

Approximating the Fisher information The amount of information about model structure or parameters which can be gained from measurements may be highly dependent on the experimental setup that is chosen [24, 25, 26, 27, 28]. Carefully planning an experiment reduces experimental effort and resources and may even allow one to answer questions which cannot be answered from unplanned experiments.

Predicting the information about a vector of unknown model parameters $\theta = [\theta_1 \cdots \theta_N]^T$ that an experimental setup can supply requires computation of the Fisher information matrix $I(\theta)$, whose elements are given by

$$(I(\theta))_{i,j} = \mathbb{E}\left[\left(\frac{\partial}{\partial\theta_i}\log f(Y;\theta)\right)\left(\frac{\partial}{\partial\theta_j}\log f(Y;\theta)\right)\right],$$

where Y is the random variable which is experimentally measured and $f(Y;\theta)$ is its distribution. In stochastic kinetic models the parameter vector θ typically contains reaction rates and, if extrinsic variability is present, extrinsic parameters such as moments of the extrinsic distribution [3] or parameters describing the fluctuations of the extrinsic variables. Population measurements, such as those provided by flow cytometry, can be viewed as a large number of independent samples $Y_1, \ldots, Y_n$ which are drawn from a marginal distribution $f(Y;\theta)$ of the underlying process. The computation of $f(Y;\theta)$, and therefore also the computation of the Fisher information matrix, requires the solution of the CME. As this is usually not available, methods which can be used to approximate the Fisher information matrix using Monte Carlo simulations have been proposed [29, 30]. Such approaches are, however, computationally expensive and hardly feasible in most applications.

An alternative approach resorts to a Gaussian assumption on the underlying Markov process [31, 32]. Under this assumption, the sample mean and variance of the measured population form a jointly sufficient statistic. Hence, computation of the Fisher information reduces to solving the differential equations that describe the dynamics of mean and variance of $f(Y;\theta)$ and computing their partial derivatives with respect to θ. There are, however, many systems where a Gaussian assumption is not reasonable [3, 33]. In such cases the method in [31] may lead to erroneous results for several reasons. First, the computed means and variances of $f(Y;\theta)$ may be inaccurate. Second, information that can be gained from higher order moments is neglected. And third, the Gaussian approximation implicitly assumes that sample mean and variance provide independent pieces of information, an assumption which is violated for all non-Gaussian distributions [34]. For instance, for a Poisson distribution the sample mean is already a sufficient statistic on its own and the variance adds no new information (see Supporting Information Section S.1.2).

In situations where a Gaussian assumption is not applicable it may still be possible to approximate the information which is provided by sample mean and variance. If the sample size is sufficiently large, the central limit theorem implies that sample mean and variance are approximately jointly Gaussian. For simplicity, assume that there is only one unknown parameter θ and that only one species is measured (a more general case is treated in the Supporting Information Section S.1.3). The information given by the mean $I_m(\theta)$ and the joint information given by mean and variance $I_J(\theta)$ can then be approximated using the special form of the Fisher information for multivariate Gaussian random variables (see Supporting Information Sections S.1.1 and S.1.3), which results in

$$I_m(\theta) \approx \tilde{I}_m(\theta) = n\frac{\left(\frac{\partial \mu_1}{\partial \theta}\right)^2}{\mu_2}, \tag{3}$$

$$I_J(\theta) \approx \tilde{I}_J(\theta) = \tilde{I}_m(\theta) + n\frac{\left(\mu_2\frac{\partial \mu_2}{\partial \theta} - \frac{\partial \mu_1}{\partial \theta}\mu_3\right)^2}{\mu_2^2\left(\mu_4 - \mu_2^2\right) - \mu_2\mu_3^2}, \tag{4}$$

where n is the size of the sample, μ_1 denotes the mean and $\mu_k, k = 2, ..., 4$ the central moments up to order 4.

These formulas are valid for any distribution which satisfies the requirements of the central limit theorem. Further, it can be shown [35] that $\tilde{I}_m(\theta)$ and $\tilde{I}_J(\theta)$ provide lower bounds on the information of the whole sample. For a Gaussian distribution, since $\mu_3 = 0$ and $\mu_4 = 3\mu_2^2$, $\tilde{I}_J(\theta)$ reduces to the correct expression for the complete information. For a Poisson distribution, since $\mu_1 = \mu_2 = \mu_3$, $\tilde{I}_J(\theta)$ reduces to $\tilde{I}_m(\theta)$, which gives the correct expression for the complete information.

Designing optimal experiments The goal of experimental design is to find the experiment which is optimal according to some criterion reflecting information about the unknown parameters. The most frequently used criteria are D-optimality, A-optimality and E-optimality which correspond to maximizing the determinant, minimizing the trace of the inverse and maximizing the minimal eigenvalue of the Fisher information matrix, respectively. In biological applications, especially in models which include extrinsic and intrinsic variability, it may be desirable to design experiments which are targeted to specific parameters or to subsets of the parameter set. For instance one might want to estimate the kinetic parameters of the model as well as possible, despite the presence of extrinsic noise. Or conversely, one may be more interested in understanding the extrinsic variability only. An optimality criterion targeted to such questions is D_s-optimality [36, 37]. It is based on partitioning the parameter vector $\theta = [\theta_1\ \theta_2]^T$ in parameters of interest θ_1 and nuisance parameters θ_2. The experiment, which allows one to obtain the confidence region with minimum volume for θ_1 can then be found by maximizing the determinant of

$$I_s(\theta) = I_{11}(\theta) - I_{12}(\theta)I_{22}(\theta)^{-1}I_{21}(\theta), \text{ where}$$
$$I(\theta) = \left[\begin{array}{cc} I_{11}(\theta) & I_{12}(\theta) \\ I_{21}(\theta) & I_{22}(\theta) \end{array}\right], \tag{5}$$

$I_{11}(\theta)$ and $I_{22}(\theta)$ are the information matrices for θ_1 and θ_2, respectively, and $I_{12}(\theta)$ and $I_{21}(\theta)$ give the cross terms between θ_1 and θ_2.

The computation of $I_s(\theta)$ requires knowledge of the true parameters θ, which are not available. This difficulty can, for instance, be overcome by evaluating the information at some initial estimate $\hat{\theta}$. If, however, this initial estimate differs significantly from the true parameter vector, the resulting experiment may be far from optimal. This is especially important for biological applications where initial estimates, if available at all, usually involve large uncertainties. Here we chose an approach which includes the uncertainty of the initial estimate in the form of a prior distribution $\pi(\theta)$ and computes the expected information with respect to $\pi(\theta)$ [38, 39] (for an overview of other methods see Supporting Information Section S.4). The corresponding optimal experiment e^* can then be obtained by solving the following optimization problem:

$$e^* = \underset{e\in\mathcal{E}}{\operatorname{argmax}}\,\{\mathbb{E}_\theta\left[\det I_s(\theta, e)\right]\}, \tag{6}$$

where the expectation is taken over $\theta \sim \pi(\theta)$, $I_s(\theta, e)$ is the information matrix for experiment e evaluated at θ and $\mathcal{E}$ is the space of possible experiments. We can now state a procedure for designing optimal experiments for the estimation of parameters of stochastic kinetic models from measurements of a heterogeneous cell population:

The optimal experimental design procedure

- Include extrinsic variability by taking reaction rates as stochastic processes as in Eq.[1].
- Derive the differential equations for the required moments using Eq.[2]. If the moment equations are non-closed and cannot be solved exactly use an approximation method [21, 22].
- Solve the differential equations to compute the moments and their partial derivatives with respect to the parameter vector as functions of θ and t.
- Choose a vector θ_1 of parameters of interest and specify the distribution $\pi(\theta)$ according to prior uncertainty about θ.
- Solve the optimization problem Eq.[6], where the total information $I_s(\theta, e)$ is replaced by the approximated information of sample mean and variance according to Eq.[4].

Some comments on practical applicability of this procedure are given in the SI Appendix Section S.1.4.

This optimal experimental design procedure can be performed in iterations with experiments. Starting from some prior distribution $\pi(\theta)$, the computations lead to optimal experiments that yield data which can be used in a parameter inference scheme [3] to compute posterior distributions. These can then in turn serve as new prior distributions for the computation of a new optimal experiment. This can be continued until the uncertainty about the parameters has been sufficiently reduced.

Results

In silico study of a gene expression system We demonstrate the proposed experimental design framework on a simple example of gene expression. The model we consider consists of the two species mRNA (M) and protein (P) and the four reactions

$$\begin{array}{ccc}
\emptyset & \xrightarrow{a_t \cdot g(t)} & M \\
M & \xrightarrow{b} & \emptyset \\
M & \xrightarrow{c} & M + P \\
P & \xrightarrow{d} & \emptyset,
\end{array}$$

where $b = 0.03, c = 0.5, d = 0.04$ and a_t is varying according to a stationary stochastic process of the form Eq.[1]. The dynamics of the moments of order up to two of M and P are then completely determined by the mean reversion speed r and mean μ_a and variance V_a of the stationary distribution of a_t (see Supporting Information Section S.2.1). Here we assume that the values of these parameters are $\mu_a = 0.5$, $V_a = 0.1$ and $r = 0.005$. We further assume that the gene can be switched between an on state (where $g(t) = 1$) and an off state (where $g(t) = 0$) using some external input, for example by adding different nutrients in nutrient shift experiments [40], by adding salt to induce the osmotic stress response [41], or with light pulses [19]. Further, throughout this section we assume that it is known that no molecules are present at time $t = 0$ (loosely speaking the gene has been off for some time at the start of the experiment) and that the degradation rates b and d are known, whereas μ_a, V_a, r and c have to be determined from the measurements. Further, for simplicity, all the computations of this section are performed locally at the "true" parameter values and prior uncertainty about the parameters is not included.

In the following we compare four experimental methods in terms of the information they can provide about the unknown parameters. For all methods we assume that the experiments are limited to a time length of $t = 300$ time units and that at most ten measurements of the protein distribution are taken during that time. The first two methods we consider are standard time course experiments, where the gene is switched on only at time zero and measurements are taken in regular time intervals (every 30 time units). In the first method (referred to as unplanned experiments) we assume that a single reporter protein is measured, whereas in the second method (unplanned dual reporter experiments) an identical copy of the gene is added to the cells, such that a second reporter protein which is conditionally independent of P, given the history of a_t, can be measured [14, 42, 43]. These two methods are compared to more sophisticated experiments where informative gene switching patterns and measurement times were identified using our experimental design framework. Thereby, the searches for the most informative experiments were performed using a Markov chain Monte Carlo-like algorithm, where we first searched for the optimal gene switching pattern for equally spaced measurement times and then sequentially placed the measurement times where they yielded maximal information for the previously found gene switching pattern. More details on this algorithm are given in the Supporting Information Section S.2.3. Again, we consider both single and dual reporter experiments (referred to as optimal perturbation and optimal dual reporter experiments, respectively). Figure 1 gives the resulting optimal perturbation experiments targeted at the parameters r and V_a (results for the remaining parameters and results for the optimal dual reporter experiments are given in Figures S.2 and S.3 in the Supporting Information). It can be seen that different perturbations and measurement times are optimal for identifying different parameters.

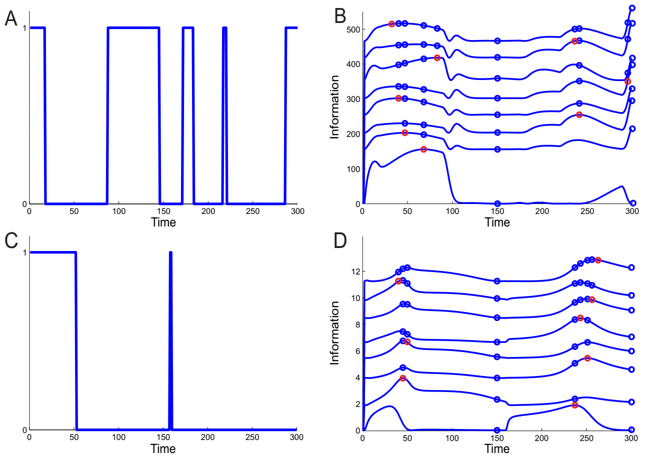

Figure 1: Optimal perturbation experiments for the parameters r (A and B) and V_a (C and D). (A,C) Gene switching patterns found by employing the experimental design procedure. A value of one corresponds to the gene being switched on, a value of zero to the gene being switched off. (B,D) Measurement times which attain the highest information for the gene switching patterns of A and C, respectively. Information is normalized by the sample size n. In both panels the first two measurement times were placed at $t = 150$ and $t = 300$ (blue circles in the bottommost curve). The bottommost curve shows the information which would be obtained by placing a third measurement at different time points between t=0 and t=300. The measurement is then placed at the time point where the information is maximal (indicated by the red circle on the bottommost curve) and the procedure is repeated for the next measurement. From bottom to top measurements are being placed one at a time (red circle on the corresponding curve) and the total information increases. The blue circles show the location of the measurements that were already placed.

The results of the comparison of the four methods are summarized in Table 1. From the first column we see that the information which can be gained from unplanned experiments is very small. This indicates that the param-

Table 1: Comparison of different experimental approaches. Information normalized by the sample size n. Rows: Information for different parameters of interest. Columns: Information which can be gained by different experimental approaches. Computations corresponding to unplanned experiments were performed with the gene being switched on only once at time zero and equally spaced measurement times. Optimal experiments include optimal gene switching patterns and sequentially placed measurement times (see Supporting Information Section S.2.3).

	Unplanned experiment	Unplanned dual reporter	Optimal perturbations	Optimal dual reporter
μ_a	0.0037	10.3106	4.4499	10.8805
V_a	0.0185	18.6882	12.8646	36.9019
c	0.0037	11.3211	4.7267	12.4333
r	48.5647	271.5223	515.6095	975.4253

eters may be practically unidentifiable. Unplanned dual reporter experiments, on the other hand, lead to much more information (second column in Table 1) and appear to be suitable for identifying both the extrinsic and the intrinsic parameters of the system. The information of the optimal perturbation experiments is given in the third column of Table 1. It can be seen that, compared to dual reporter experiments, more information is obtained for the parameter r, whereas dual reporter experiments lead to more information for μ_a, V_a and c. Hence, depending on the objective of the study, different experimental strategies are preferable. The fourth experimental method, which combines optimal perturbations and dual reporters, naturally leads to the most information for all parameters (fourth column in Table 1). Note, however, that the increase in information compared to unplanned dual reporter experiments is very small if μ_a or c are the parameters of interest, which indicates that additionally perturbing the system may not be worth the effort.

Finally, we also computed the information for the first two methods under a Gaussian assumption. Our results (Supporting Information Table S.1) show that for many objectives a Gaussian assumption leads to information estimates which are overly optimistic. This is most likely due to the independence assumption of sample mean and variance which is implicitly imposed by a Gaussian approximation.

Characterizing extrinsic noise in a light-induced gene expression system Next we study a light-switch gene expression system which has been engineered in yeast [19]. The authors used a light responsive module to initiate transcription by shining red light on the yeast culture and to terminate transcription by shining far-red light. They then proposed a control scheme to regulate the mean amount of protein in the population. The development of more sophisticated control schemes (for example, to allow one to also control the protein variance) requires a proper characterization of the sources of variability in the system. To this end our framework can serve to compare the utility of different experiments and ultimately to design the experiments which are optimal for the characterization of the different noise sources. We thus developed a stochastic version of the model in [19] with time varying extrinsic variability. This introduced four additional parameters which were not identified in [19]. To characterize uncertainty about these parameters we chose independent uniform prior distributions and computed the expectation of the information which is provided by the experiments reported in Figure 1 of [19] about each of the additional parameters according to Eq.[5]. Thereby, the remaining parameters which were already identified in [19] were fixed to their known values (see Supporting Information Section S.3.1 and Section S.3.2). The results are shown in Tables S.2 and S.3 in the Supporting Information. Which experiment is best again depends on the objective. For instance, the experiment where a red light pulse is applied at the beginning and a far-red light pulse after 30 minutes is best to identify the protein production rate but worst for all other parameters. This is most likely due to the fact that if the gene is switched off, the mRNA production rate is set to zero and the parameters describing this rate do not influence the dynamics anymore.

Further, our results show that even though the experiments in Figure 1c of [19] were performed over a longer time and contain more measurements than the experiments in Figure 1e of [19], the experiments in Figure 1e provide much more information about the mean reversion speed r of the extrinsic fluctuations. This suggests that experiments where the gene is expressed for short time intervals with silent periods in between could allow one to determine r and thus to test whether extrinsic variability can be assumed to be constant or whether a stochastic process description as used in this paper is required. Our results in Section S.3.3 in the Supporting Information indicate that indeed, contrary to other experiments, the variance measured in the experiments in Figure 1e in [19] cannot be explained very well by a model with constant extrinsic variability.

Finally, we also employed our experimental design framework to search for experiments which carry high information about r. The light-pulse pattern and measurement times we determined with our method are shown in Figure 2. They lead to an experiment which carries close to four times more information than any of the experiments in [19] suggesting that our experimental design framework can be a valuable tool for characterizing variability in this system.

Discussion

While knowledge about biological mechanisms is constantly growing, our understanding of the stochasticity of biological systems and its influence on system dynamics remains rather limited. Many different sources of variability may play a role and neglecting any one of them may lead to substantial errors. For instance, neglecting extrinsic variability in the gene expression model and us-

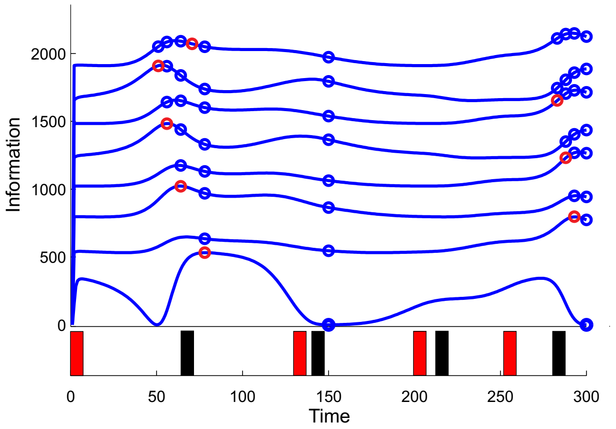

Figure 2: Results of experimental design targeted at the parameter r for the light-switch gene expression system. The optimal light-pulse pattern is shown at the bottom of the figure. Red rectangles correspond to red light pulses, black rectangles correspond to far-red light pulses. The upper part of the figure shows the optimal measurement times, where the interpretation is the same as in Figure 1.

ing the ideas in [44] to quantify protein burst size and frequency from measurements of the protein distribution would lead to conclusions which are more than an order of magnitude wrong (see Supporting Information Section S.2.2). Allowing reaction rates to vary between individuals in a cell population, either by taking them as unknown constant random variables or by allowing them to be governed by stochastic differential equations, offers a way to incorporate extrinsic variability in a model and enables model-based studies of heterogeneous cell populations. We showed how the information about unknown parameters of such models which is provided by means and variances of measured populations can be approximated by solving a set of ordinary differential equations. The derived formulas apply to any kind of system with the only assumption that the measured population is of sufficient size for the application of the central limit theorem. This opens up the possibility to pose many interesting questions: do the measurements contain enough information to separate the different noise sources? How much information can be gained by measuring the variance in addition to the mean? And most importantly: what is the most informative experiment? By means of examples we demonstrated that unplanned experiments may not contain enough information to separate the different sources of variability and that designing experiments based on intuition alone may not be sufficient. For instance, placing all the measurements either very early or very late in the experiment turns out to be optimal for identifying the mean reversion speed in our examples (see Figures 1 and 2), but appears very unintuitive at a first glance.

Our results (Table 1 and Supporting Information Figure S.2) show that the optimal experiments are highly dependent on the chosen objective. In some cases introducing a dual reporter yields high information, in other cases perturbing the system with input stimuli is preferable. A study of the system using the experimental design framework presented in this paper allows a comparison of the different experimental approaches and enables one to choose the approach which is most likely to be successful for the given objective. The resulting experiments can in turn be used to refine the model and to update the parameter estimates, giving rise to an iterative procedure of successive rounds of computations and experiments.

In the light-switch gene expression system the computation of the information contents of the different experiments shows that perturbing the system with different light pulse sequences can highlight different features of the system. This suggests that well chosen gene induction patterns may allow one to uncover features of the system which remain hidden in unplanned experiments. For instance Figure S.4 in the Supporting Information suggests that temporal fluctuations in the extrinsic variable may play a role for this system. Perturbing a system with light pulses to understand the variability may seem to be limited to this specific engineered system. However, a similar strategy could also be employed by exploiting naturally occurring biological mechanisms. For example, in [3] the authors studied gene expression in yeast in response to osmotic pressure. Different salt concentrations lead to different residence times of the signaling molecule Hog1 in the nucleus and thereby created different input signals to the downstream gene expression system. In that system multiple subsequent salt stresses, which can for instance be implemented using a micro-fluidic device as in [41], could serve as the equivalent of the multiple light-pulses used in this paper and might give further insights into the specific nature of the system.

Acknowledgements The work was supported in part by the European Commission under the project MoVeS.

Author Contributions J.R. and J.L. designed research. J.R. developed the method and performed the computations. A.M. assisted with supplementary calculations. J.R. and J.L. wrote the paper.

Supporting Information A PDF document with supporting discussions and results is available at http://control.ee.ethz.ch/∼jakorues/DesignSI/.

References

[1] Munsky B, Trinh B, Khammash M. 2009 Listening to the noise: random fluctuations reveal gene network parameters. *Mol Syst Biol* 5:318.

[2] Neuert F, Munsky B, Tan RZ, Teytelman L, Khammash M, van Oudenaarden A. 2013 System identification of signal-activated stochastic gene regulation. *Science* 339:584–587.

[3] Zechner C, Ruess J, Krenn P, Pelet S, Peter M, Lygeros J, Koeppl H. 2012 Moment-based inference predicts bimodality in transient gene expression. *Proc Natl Acad Sci USA* 109:8340–8345.

[4] Singh A, Razooky B, Dar R, Weinberger L. 2012 Dynamics of protein noise can distinguish between alternate sources of gene-expression variability. *Mol Syst Biol* 8:607.

[5] Gillespie D. 1992 A rigorous derivation of the chemical master equation. *Physica A* 188:404–425.

[6] Munsky B, Khammash M. 2006 The finite state projection algorithm for the solution of the chemical master equation. *J Chem Phys* 124:044104.

[7] van Kampen N. 2006 *Stochastic Processes in Physics and Chemistry* (Elsevier Science, Amsterdam).

[8] Gillespie D. 2000 The chemical Langevin equation. *J Chem Phys* 113:297–306.

[9] Snijder B, Pelkmans L. 2011 Origins of regulated cell-to-cell variability. *Nat Rev Mol Cell Biol* 12:119–125.

[10] Shahrezaei V, Ollivier J, Swain P. 2008 Colored extrinsic fluctuations and stochastic gene expression. *Mol Syst Biol* 4:196.

[11] Hilfinger A, Paulsson J. 2011 Separating intrinsic from extrinsic fluctuations in dynamic biological systems. *Proc Natl Acad Sci USA* 108:12167–12172.

[12] Koeppl H, Zechner C, Ganguly A, Pelet S, Peter M. 2011 Accounting for extrinsic variability in the estimation of stochastic rate constants. *Int J Robust Nonlin* 22:1103–1119.

[13] Hilfinger A, Chen M, Paulsson J. 2012 Using temporal correlations and full distributions to separate intrinsic and extrinsic fluctuations in biological systems. *Phys Rev Lett* 109:248104.

[14] Elowitz M, Levine A, Siggia E, Swain P. 2002 Stochastic gene expression in a single cell. *Science* 297:1183–1186.

[15] Colman-Lerner A, Gordon A, Serra E, Chin T, Resnekov O, Endy D, Pesce G, Brent R. 2005 Regulated cell-to-cell variation in a cell-fate decision system. *Nature* 437:699–706.

[16] Volfson D, Marciniak J, Blake W, Ostroff N, Tsimring L, Hasty J. 2005 Origins of extrinsic variability in eukaryotic gene expression. *Nature* 439:861–864.

[17] Chabot J, Pedraza J, Luitel P, Van Oudenaarden A. 2007 Stochastic gene expression out-of-steady-state in the cyanobacterial circadian clock. *Nature* 450:1249–1252.

[18] Rosenfeld N, Young J, Alon U, Swain P, Elowitz M. 2005 Gene regulation at the single-cell level. *Science* 307:1962–1965.

[19] Milias-Argeitis A, Summers S, Stewart-Ornstein J, Zuleta I, Pincus D, El-Samad H, Khammash M, Lygeros J. 2011 In silico feedback for in vivo regulation of a gene expression circuit. *Nat Biotechnol* 29:1114–1116.

[20] Hespanha J. 2006 Modeling and analysis of stochastic hybrid systems. *IEE Proceedings Control Theory And Applications* 153:520–535.

[21] Ruess J, Milias-Argeitis A, Summers S, Lygeros J. 2011 Moment estimation for chemically reacting systems by extended Kalman filtering. *J Chem Phys* 135:165102.

[22] Singh A, Hespanha J. 2011 Approximate moment dynamics for chemically reacting systems. *IEEE Trans Automat Contr* 56:414–418.

[23] Ale A, Kirk P, Stumpf M. 2013 A general moment expansion method for stochastic kinetic models. *ArXiv e-prints. 1303.5848. q-bio.MN.*

[24] Bandara S, Schlöder J, Eils R, Bock H, Meyer T. 2009 Optimal experimental design for parameter estimation of a cell signaling model. *PLoS Comput Biol* 5:e1000558.

[25] Busetto A, Ong C, Buhmann J. 2009 Optimized expected information gain for nonlinear dynamical systems. *Proceedings of the 26th Annual International Conference on Machine Learning* pp 97–104.

[26] Franceschini G, Macchietto S. 2008 Model-based design of experiments for parameter precision: State of the art. *Chem Eng Sci* 63:4846–4872.

[27] Liepe J, Filippi S, Komorowski M, Stumpf M. 2013 Maximizing the information content of experiments in systems biology. *PLoS Comput Biol* 9:e1002888.

[28] Zechner C, Nandy P, Unger M, Koeppl H. 2012 Optimal Variational Perturbations for the Inference of Stochastic Reaction Dynamics. *IEEE 51st Annual Conference on Decision and Control (CDC)* pp 5336–5341.

[29] Rathinam M, Sheppard P, Khammash M. 2010 Efficient computation of parameter sensitivities of discrete stochastic chemical reaction networks. *J Chem Phys* 132:034103.

[30] Gunawan R, Cao Y, Petzhold L, Doyle F. 2005 Sensitivity analysis of discrete stochastic systems. *Biophys J* 88:2530–2540.

[31] Komorowski M, Costa M, Rand D, Stumpf M. 2011 Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc Natl Acad Sci USA* 108:8645–8650.

[32] Wlodarczyk M, Lipniacki T, Komorowski M. 2013 Functional redundancy in the NF-κB signalling pathway. *ArXiv e-prints. 1303.3109. q-bio.QM.*

[33] Arkin A, Ross J, McAdams H. 1988 Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells. *Genetics* 149:1633–1648.

[34] Lukacs E. 1942 A characterization of the normal distribution. *The Annals of Mathematical Statistics* 13:91–93.

[35] Jarret R. 1984 Bounds and expansions for Fisher information when the moments are known. *Biometrika* 71:101–113.

[36] Hunter W, Hill W, Henson T. 1969 Designing experiments for precise estimation of all or some of the constants in a mechanistic model. *Can J Chem Eng* 47:76–80.

[37] Walter E, Pronzato L. 1990 Qualitative and quantitative experiment design for phenomenological models - a survey. *Automatica* 26:195–213.

[38] Pronzato L, Walter E. 1985 Robust experimental design via stochastic approximation. *Math Biosci* 75:103–120.

[39] Chaloner K, Larntz K. 1989 Optimal Bayesian design applied to logistic regression experiments. *J Stat Plan Inference* 21:191–208.

[40] Menolascina F, di Bernardo M, di Bernardo D. 2011 Analysis, design and implementation of a novel scheme for in-vivo control of synthetic gene regulatory networks. *Automatica* 47:1265–1270.

[41] Uhlendorf J, Miermont A, Delaveau T, Charvin G, Fages F, Bottani S, Batt G, Hersen P. 2012 Long-term model predictive control of gene expression at the population and single-cell levels. *Proc Nat Acad Sci USA* 109:14271–14276.

[42] Swain P, Elowitz M, Siggia E. 2002 Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* 99:12795–12800.

[43] Bowsher C, Swain P. 2012 Identifying sources of variation and the flow of information in biochemical networks. *Proc Nat Acad Sci USA* 109:E1320–E1328.

[44] Friedman N, Cai L, Xie S. 2006 Linking stochastic dynamics to population distribution: an analytical framework of gene expression. *Phys Rev Lett* 97:168302.